Simplifying and Optimizing Your Data Transfers

June 15th, 2016

NASA Advanced Supercomputing Division

Overview

- Data transfers using Shift
  - What is Shift?
  - Why should you use it?
- Basic transfer tasks
  - Copying files
  - Synchronizing directories
  - Creating/extracting tar files
- Optimizing transfer performance
  - Shift tuning options and when you should use them
- Common mistakes/misunderstandings
  - What they are and how to avoid them

Question? Use the Webex chat facility to ask the Host

Self-Healing Independent File Transfer (Shift)

- What is it?
  - Automated file transfer mechanism
  - Supports local, intra-enclave, and remote transfers
  - Has facilitated transfers of > 40 PB of data since April 2012
- Why should you use it?
  - Unified interface with simple cp/scp syntax
  - Fire and forget transfers
    - Takes care of numerous details so you don't have to
    - Uses best practices by default
  - Advanced performance and reliability features
  - Saves you time and effort
    - Reduces transfer time
    - Reduces learning curve
    - Reduces manual transfer management

Question? Use the Webex chat facility to ask the Host

Basic Transfer Tasks

Shift's Three Main Capabilities

- Copying files
- Synchronizing directories
- Creating/extracting tar files

Question? Use the Webex chat facility to ask the Host

Copying Files With Traditional Tools

- Local transfers
  - **pfe%** cp /file1 /file2
- Intra-enclave transfers
  - **pfe%** scp /file1 lfe:/file2
- Remote transfers
  - **yourhost%** sup scp /file1 pfe:/file2

Question? Use the Webex chat facility to ask the Host

Secure Unattended Proxy (SUP) Review

- What is it?
  - Mechanism that allows specific commands to be invoked on HEC front-ends from external systems without SecurID for up to a week
- Usage
  - Download "sup" (http://www.nas.nasa.gov/hecc/support/kb/file/9)
  - Make client executable (chmod 700 sup)
  - Move client to directory in your $PATH (mv sup ~/bin)
  - Authorize directories for writes on both pfes and lfes
    - **pfe%** echo /u/user > ~/.meshrc;
    - **pfe%** echo /nobackup/user >> ~/.meshrc
    - **pfe%** scp ~/.meshrc lfe:
  - Prepend "sup" to normal commands (sup scp file pfe:/nobackup/user)
- More information
  - http://www.nas.nasa.gov/hecc/support/kb/entry/145

Question? Use the Webex chat facility to ask the Host

Copying Files With Shift

- Local transfers (just like "cp")
  - **pfe%** cp /file1 /file2
  - **pfe%** shiftc /file1 /file2
- Intra-enclave transfers (just like "scp")
  - **pfe%** scp /file1 lfe:/file2
  - **pfe%** shiftc /file1 lfe:/file2
- Remote transfers (just like "sup scp")
  - **yourhost%** sup scp /file1 pfe:/file2
  - **yourhost%** sup shiftc /file1 pfe:/file2
    - shiftc is embedded in sup client
    - All examples with "shiftc" can be done with "sup shiftc"

Question? Use the Webex chat facility to ask the Host

Common Copy Options

- Recursive transfers (-r/-R/--recursive...just like cp/scp)
  - Copy directories recursively
- Symbolic link dereferencing
  - Never follow links (-P/--no-dereference...just like cp)
  - Always follow links (-L/--dereference...just like cp)
- Directory handling
  - Create missing parent directories (-d/--directory...just like install)
  - Treat target as a file (-T/--no-target-directory...just like cp/install)
- Secure remote transfers (--secure) (decreases performance)
  - Only use remote transports that encrypt data streams
  - Use more secure ciphers and MAC algorithms for ssh connections
- Feature disablement
  - Do not send status emails (--no-mail)
  - Do no preserve times, mode, ownership, xattrs, acls (--no-preserve)

Question? Use the Webex chat facility to ask the Host

After Starting A Transfer

- Shift prints transfer id
  - "Shift id is 1"
  - The id can be used to manage/monitor a particular transfer
- Shift detaches and begins the transfer
  - "Detaching process (use --status option to monitor progress)"
  - You do not need to stay logged on to the origin system
  - You will be notified by email of completion, errors, and/or warnings (unless --wait or --no-mail specified)

Stopping/Restarting/Monitoring Transfers

- A running transfer may be stopped at any time from any host
  - shiftc --stop --id=N
  - Batches of file operations in progress will run to completion
- A stopped/failed transfer (state = stop/error) may be restarted
  - shiftc --restart --id=N
  - Completed operations will not be run again
  - Failed/unattempted operations will be retried/attempted
  - This is the best and fastest way to recover from transient errors
  - Must restart on original host or one with equivalent file system access
- Shift provides status and history of transfers
  - shiftc --status
  - shiftc --history
  - Transfer data only kept for one week after completion/error/stop

Transfer Status (--status)

```
pfe% shiftc --status
```

```
id | state | dirs | files |       file size |  date |  run |      rate
   |       | sums | attrs |        sum size |  time | left |
---+-------+------+-------+-----------------+-------+------+---------
 1 | done  |  0/0 |   1/1 |       92KB/92KB | 10/03 |   2s |    46KB/s
   |       |  0/0 |   0/0 |       0.0B/0.0B | 17:06 |      |
 2 | done  |  0/0 |   1/1 |       92KB/92KB | 10/03 |   8s |  11.5KB/s
   |       |  0/0 |   1/1 |       0.0B/0.0B | 17:06 |      |
 3 | done  |  1/1 |   2/2 |       99KB/99KB | 10/03 |   1s |    99KB/s
   |       |  4/4 |   0/0 |     198KB/198KB | 17:07 |      |
 4 | error |  1/1 |   1/2 |       92KB/99KB | 10/03 |   3s |  30.7KB/s
   |       |  0/0 |   0/0 |       0.0B/0.0B | 17:08 |      |
 5 | done  |  1/1 | 64/64 |   65.5GB/65.5GB | 10/03 |  29s |  2.26GB/s
   |       |  0/0 |   0/0 |       0.0B/0.0B | 17:09 |      |
```

Question? Use the Webex chat facility to ask the Host

Detailed Transfer Status (--status --id=N)

```
yourhost% sup shiftc --status --id=2

state | op     | target                | size |  date | length |    rate
      | tool   | info                  |      |  time |        |
------+--------+-----------------------+------+-------+--------+-------
done  | cp     | lfe2:/u/user1/file1 | 92KB | 10/03 |     5s | 18KB/s
      | bbftp  | -                     |      | 17:06 |        |
done  | chattr | lfe2:/u/user1/file1 |    - | 10/03 |     1s |      -
      | sftp   | -                     |      | 17:06 |        |

yourhost% sup shiftc --status --id=4 --state=error

state | op    | target               | size |  date | length | rate
      | tool  | info                 |      |  time |        |
------+-------+----------------------+------+-------+--------+-----
error | cp    | /tmp/dir2/file2      |  7KB |     - |      - |    -
      | rsync | rsync: send_files    |      |       |        |
      |       | failed to open:      |      |       |        |
      |       | Permission denied    |      |       |        |
```

Question? Use the Webex chat facility to ask the Host

Transfer History (--history)

```
pfe% shiftc --history

    id | origin               | command
   ---+--------------------+-----------------------------------------
     1 | pfe21.nas.nasa.gov | shiftc file1 /tmp/dir1
     2 | pfe21.nas.nasa.gov | shiftc --no-verify -p file1 lfe2:
     3 | your_localhost     | sup shiftc -r /tmp/dir1 lfe2:/tmp/dir2
     4 | your_localhost     | sup shiftc -r --secure lfe2:/tmp/dir2 .
     5 | pfe21.nas.nasa.gov | shiftc -r --hosts=4 bigdir1 /nobackup/user1/bigdir2
```

Question? Use the Webex chat facility to ask the Host

Synchronizing Directories (--sync)

- Shift can synchronize directories like rsync
  - Files with same modification time and size at src and dst are skipped
  - Files that don't exist at the dst are copied
  - Files with differing time/size are checksummed and reconciled using partial transfers
- Force checksum verification using -I/--ignore-times (just like rsync)
- Target follows normal -r semantics
  - Source placed beneath target if target exists, renamed to target otherwise
  - To synchronize directory to same name, specify parent of desired destination
    - shiftc --sync -r /dir1 /dir1_parent (to sync /dir1 with /dir1_parent/dir1)
  - To synchronize to different name, use -T/--no-target-directory
    - shiftc --sync -r -T /dir1 /dir2 (to sync /dir1 with /dir2)
- Synchronizing to lou file systems can be very time-consuming
  - Files may need to be retrieved from tape to checksum

Question? Use the Webex chat facility to ask the Host

Creating/Extracting Tar Files (--create-tar/--extract-tar)

- Transfer to/from tar file instead of transfer to/from directory
  - Uses high-speed transports underneath
  - Supports all standard Shift features (e.g. verification, parallelization, ...)
  - Recursive directory traversal (-r) implied (just like tar)
  - Up to 20x GNU tar performance on a single host
- Tar creation
  - tar cf dirfile.tar /some/dir /some/file
  - shiftc --create-tar /some/dir /some/file dirfile.tar
- Tar extraction
  - tar xf dirfile.tar /some/dir
  - shiftc --extract-tar dirfile.tar /some/dir
- Mutually exclusive with --sync
  - Not practical to change sizes of files within tar archives

Question? Use the Webex chat facility to ask the Host

Remotely Creating/Extracting Tar Files

- Like normal Shift transfers, source(s) or target may be on a remote host
- Remote tar creation with local files
  - shiftc --create-tar /some/dir /some/file lfe:dirfile.tar
- Local tar creation with remote and local files
  - shiftc --create-tar lfe:/some/dir /some/file dirfile.tar
- Remote tar extraction from local files
  - shiftc --extract-tar dirfile1.tar dirfile2.tar lfe:/some/dir
- Local tar extraction with remote and local files
  - shiftc --extract-tar lfe:dirfile1.tar dirfile2.tar /some/dir

Question? Use the Webex chat facility to ask the Host

Shift Tar Options

- Splitting files over multiple tar files (--split-tar)
  - Files larger than tape size are detrimental to lou file system
  - Tar files are split by default at 500 GB
    - Can be changed using --split-tar (e.g. --split-tar=2t) (disable with 0)
  - Split tars for "name.tar" are called "name.tar-i.tar" for i ≥ 1
    - Can all be extracted at once using "--extract-tar name.*tar"
- Creating table of contents (--index-tar)
  - Large tar files on lou will eventually be migrated to tape
  - See "ls -l" output of contents without retrieving from tape
    - Creates "name.tar.toc" or "name.tar-i.tar.toc" for each split i
  - See "msum --check-tree" output of contents without retrieving from tape
    - Creates "name.tar.sum" or "name.tar-i.tar.sum" for each split i
  - May become default in the near future

More Information on Basic Usage

- Knowledge base article
  - http://www.nas.nasa.gov/hecc/support/kb/entry/300
- Man page
  - "man shiftc" on any HEC front-end

Question? Use the Webex chat facility to ask the Host

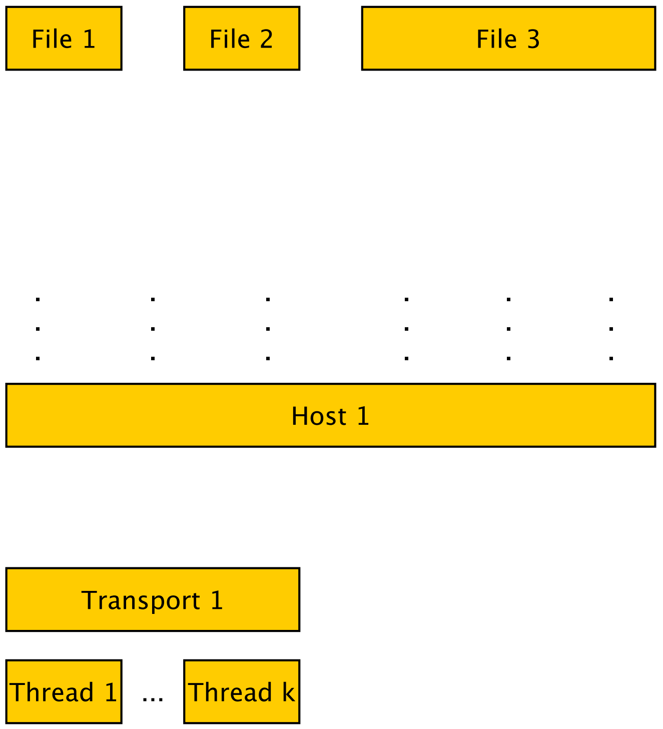
Optimizing Transfer Performance

Shift Transfer Optimization

- Shift behavior and correspondence to tuning options
- Tuning options you should care about
- Ways you can help Shift on your external system(s)

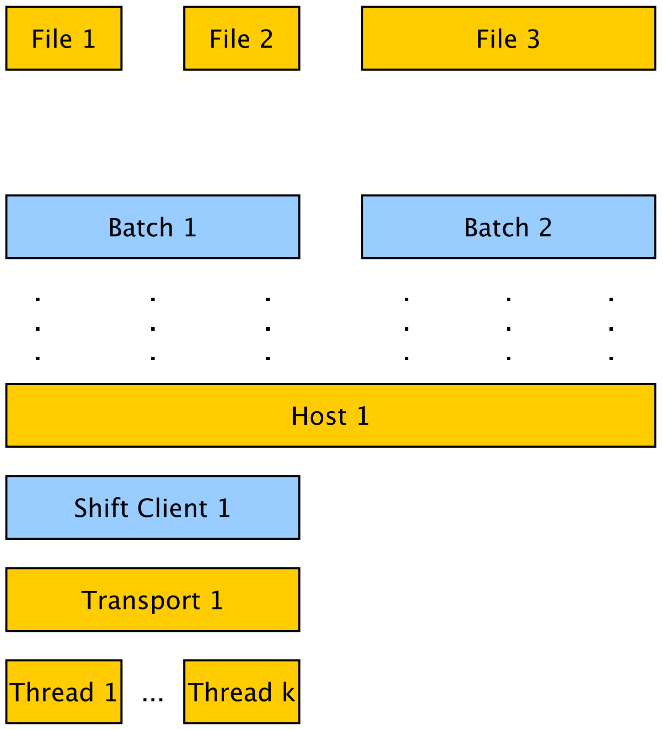
Question? Use the Webex chat facility to ask the Host

Traditional Transfers



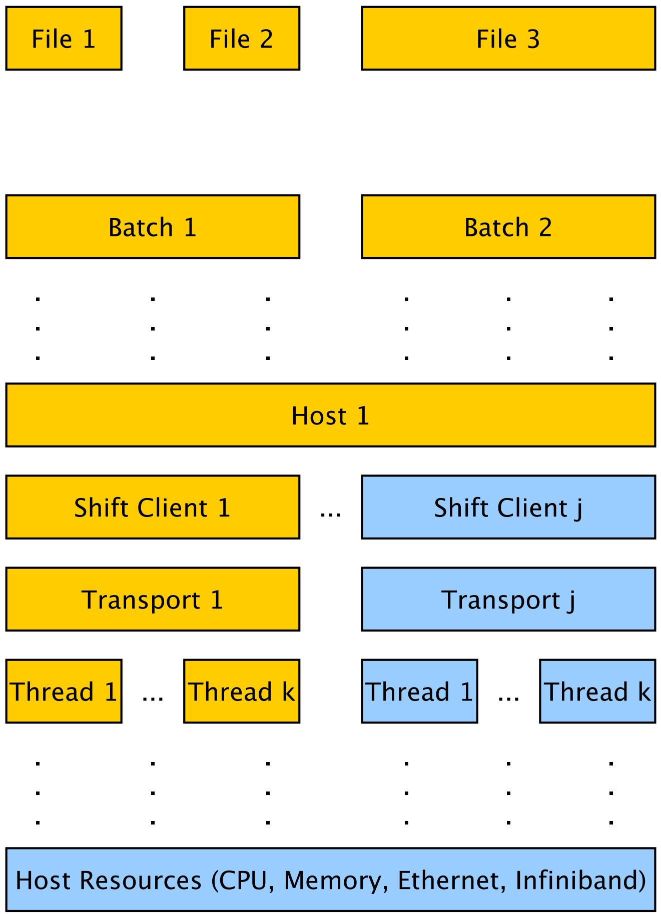
Question? Use the Webex chat facility to ask the Host

Shift Transfers



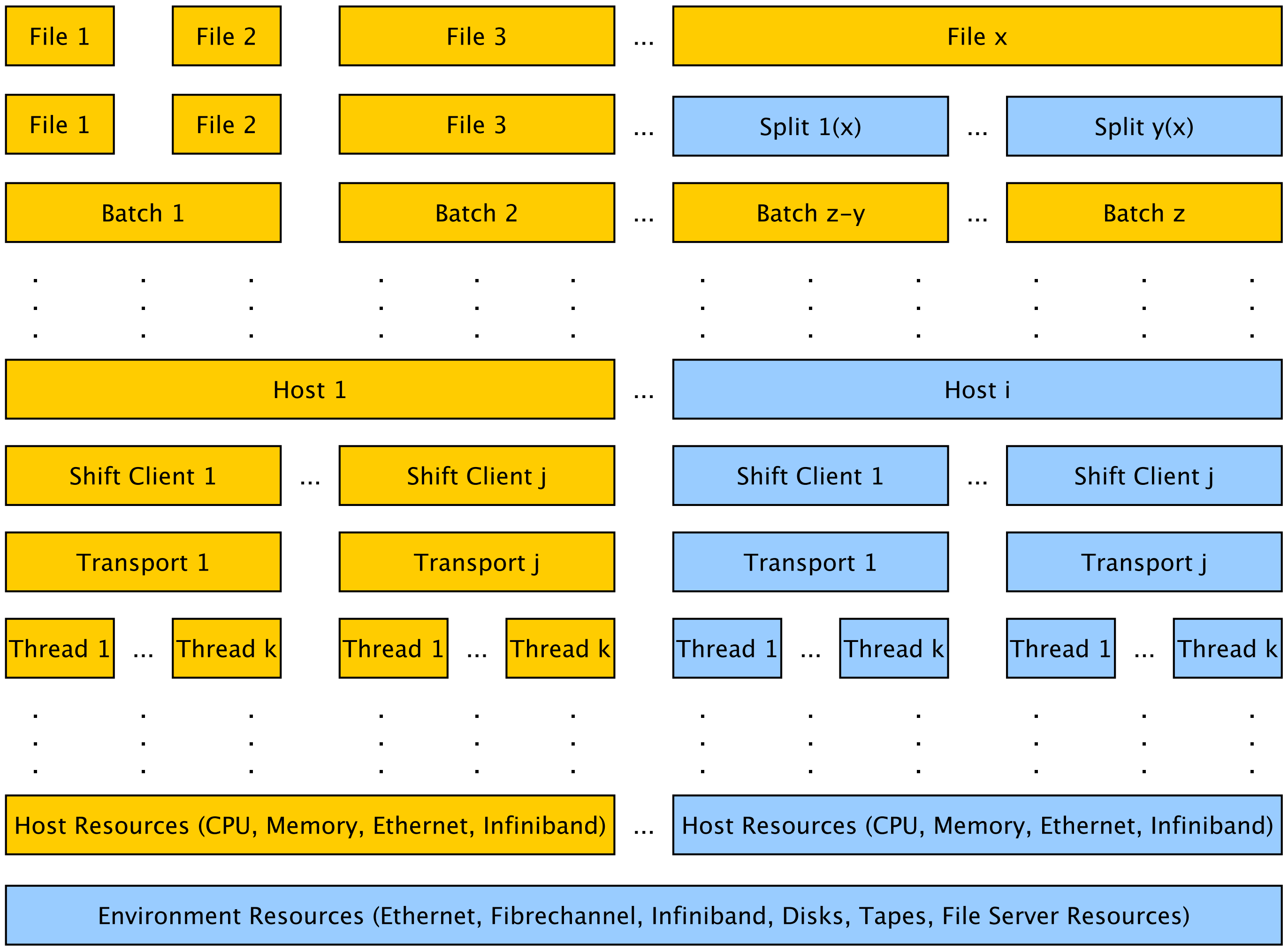
Question? Use the Webex chat facility to ask the Host

Shift Transfers With Multiple Clients



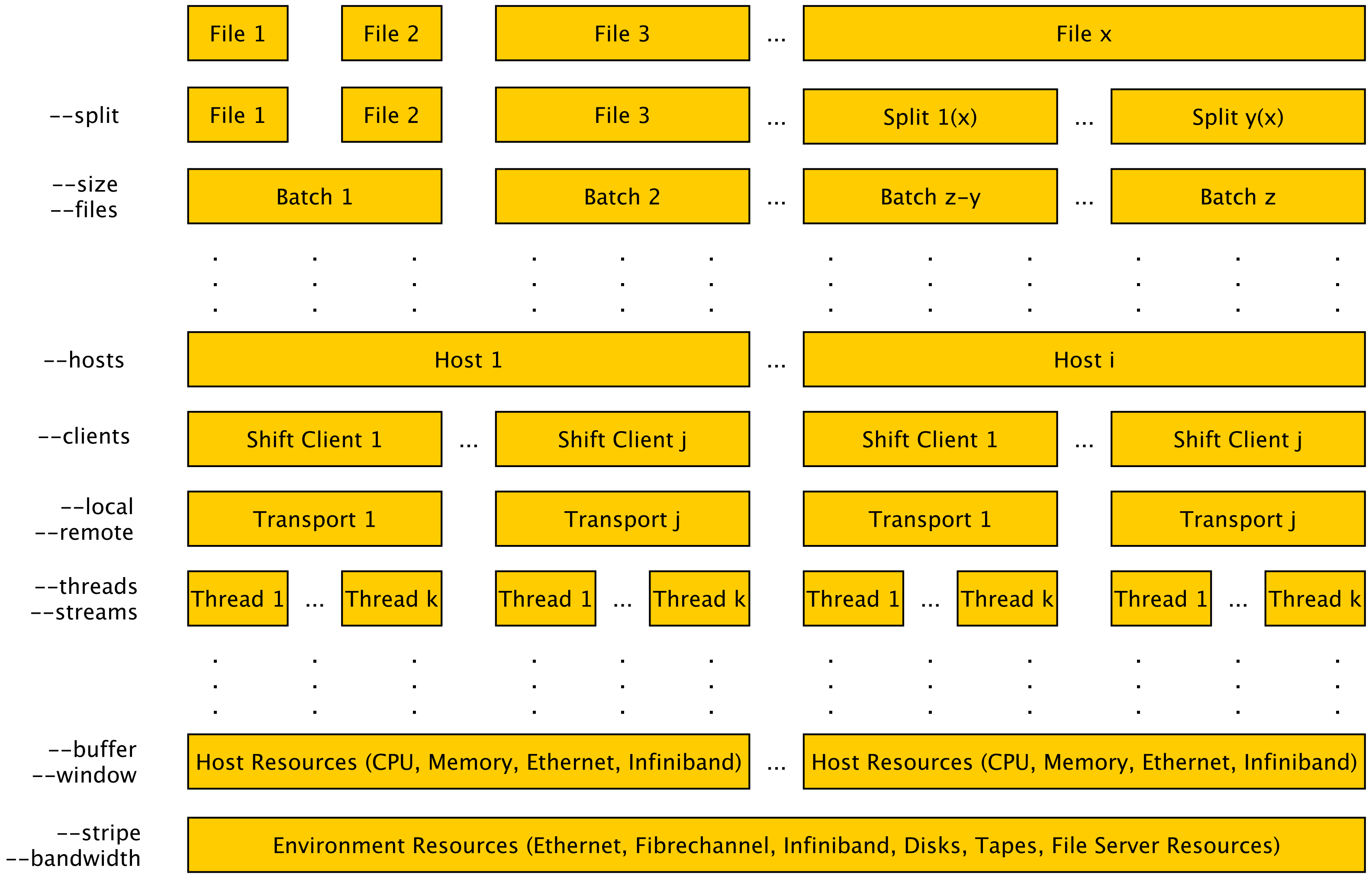
Question? Use the Webex chat facility to ask the Host

Shift Transfers With Multiple Hosts



Question? Use the Webex chat facility to ask the Host

Shift Tuning Options



Question? Use the Webex chat facility to ask the Host

That's A Lot Of Settings!

- The bad news
  - The number of settings can be overwhelming
- The good news
  - Shift takes care of most of them for you
    - Batch size set appropriately for default thread count
    - Transfers within NAS now parallelized on two hosts
    - Transports dynamically chosen to optimize each batch
    - Files striped to maximize write/future read performance
    - Buffer size set to reasonable value
    - Bandwidth approximated via heuristics
    - TCP window/streams set using bw/latency/OS limits

Question? Use the Webex chat facility to ask the Host

Multiple Hosts (--hosts=N)

- Single easiest way to improve performance
- Within NAS
  - Default now --hosts=2
  - 6 user-accessible hosts mount cxfs (i.e. lou) (lfe/lsn)
    - More than 6 hosts has negative impact for transfers to/ from lou file systems
  - 22 user-accessible hosts mount lustre ([lp]fe/bridge/[lp]sn)
    - Lustre to lustre transfers can be done at very high speed
- Outside NAS
  - Src or dst must be on file system shared by multiple hosts at your site
  - More details later

File Splitting (--split=Ng)

- Multiple clients (on same or different hosts) are only effective if there is enough work to keep them busy
  - Large files can cause imbalances in client workload
  - Splitting allows large files to be processed by different clients
- Splitting is not currently enabled by default
  - Not compatible with bbftp and rsync
  - May be enabled in future since faster options are now available for both
- The split size should be set to at least the batch size
- It is beneficial to split the largest file into several times the number of total clients
  - Allows faster clients to process more than slower clients
  - Less work is lost if a client/host goes down for some reason

Question? Use the Webex chat facility to ask the Host

Multiple Clients On A Single Host (--clients=N)

- Multiple clients were more useful before latest update
  - All built-in transports are now multi-threaded
- Two transports benefit most from this
  - Rsync is single threaded so multiple clients are the only way to parallelize its execution
  - Bbftp is multi-threaded but only operates on one file at a time so multiple clients allow it to parallelize across files
    - Detrimental to performance with big enough files (> 4 GB)
- Only use --clients when forcing these transports
  - --local/--remote=rsync or --local/--remote=bbftp
- All other transports perform better with increase in threads rather than clients

Question? Use the Webex chat facility to ask the Host

Transport Thread Count (--threads=N)

- The default number of threads (4) is set low on purpose
  - At higher counts, a single transfer can exhaust a host's I/O bandwidth and impact all users
  - HEC front-ends are shared by all users for many purposes
- Higher thread counts are useful in lustre to lustre transfers
  - Within jobs on dedicated compute nodes – 32 threads achieves max
  - On front-ends – 8 threads is reasonable maximum
- Resist the temptation to increase threads on lou transfers
  - Lou front-ends are limited in number and often have higher loads
  - File system capacity is limited and migration constrained by tape speed
- Higher thread counts need bigger batches to keep threads busy
  - Batch size (--size) should be at least --threads GB

Question? Use the Webex chat facility to ask the Host

Lustre Stripe Count (--stripe=N, --stripe=Ng)

- Shift automatically selects an appropriate stripe count for files transferred to lustre file systems
  - Does not work with bbftp due to use of temporary files
  - Shift preserves non-default striping when src also on lustre
- Two main reasons to touch this setting
  - You know the behavior of a parallel application that will access the data is optimized at a specific stripe count
    - Using --stripe=N will set stripe count of all files to N
    - Using --stripe=Ng will use one stripe per N GB
  - You have set the striping on the target directory and want to use default lustre striping behavior (--stripe=0)

Options That Help Shift (Remote Transfers)

- --bandwidth
  - Shift uses crude heuristics to guess your bandwidth
  - Specifying --bandwidth allows Shift to better optimize TCP window size and number of streams
    - For example, --bandwidth=10g
- --host-list
  - Shift remembers remote hosts you have invoked it on and their shared file systems for future --hosts invocations
  - You can more rapidly let Shift know about other hosts with equivalent file system access using --host-list
    - For example, --host-list=mycluster1,mycluster2,mycluster3
  - After first time, can just use --hosts

Question? Use the Webex chat facility to ask the Host

Actions That Help Shift (Remote Transfers)

- Installing supported software
  - Shift can only use bbftp if installed and in your $PATH
    - http://doc.in2p3.fr/bbftp
  - New built-in fish-tcp transport outperforms bbftp in most cases so bbftp not as important as it used to be
    - Need perl >= 5.10.1 with threads (perl -V lgrep THREAD)
- Adjusting operating system TCP window settings
  - Keep TCP pipeline full while packets in flight
  - Determines max performance per TCP stream
    - max throughput = max window / latency
  - sysctl net.core.[wr]mem_max and net.ipv4.tcp_[wr]mem
    - 100MB for 10 Gb/s, 10MB for 1 Gb/s, 1MB for 100 Mb/s

Question? Use the Webex chat facility to ask the Host

Common Mistakes/Misunderstandings

Remote Transfers

- Please install "sup" client as yourself in a writable location
  - Critical fixes are provided through built-in updates
    - Update when prompted – there was a reason for the update
    - Remote admins cannot easily track the need for updates
- Make sure ~/.meshrc exists and is configured on both the pleiades and lou home file systems
- A src/dst with "user@" currently does not do anything
  - Two options if remote/NAS user name differs
    - Use "sup -u NAS_user" on every invocation (or create shell alias)
    - Modify ~/.ssh/config (if not using NAS ssh template)
      - Host sup.nas.nasa.gov sup-key.nas.nasa.gov
        - User NAS_user
  - Will probably add better support for "user@" syntax in near future

Question? Use the Webex chat facility to ask the Host

Parallel Remote Transfers

- --hosts has no effect unless you have multiple hosts on your side with access to transfer src or dst on shared file system
  - Use --host-list option first time so Shift can learn about other hosts
- You have multiple hosts but --hosts not working
  - Check all hosts in each other's ssh known hosts
  - All hosts must have "sup" in default $PATH
  - All hosts must be accessible via publickey or hostbased authentication
    - Have ssh agent running on origin if publickey authentication used
  - Can test these conditions with single command
    - ssh -oBatchMode=yes otherhost sup
      - "Host verification failed" – bad known hosts
      - "Command not found" – bad path
      - "Permission denied" – bad authentication
      - "Usage: sup" – setup is good so some other issue if not working

Many Transfers At Once

- Avoid starting many transfers on the same host at the same time
  - Even one transfer can potentially max out a host's resources
  - Many transfers can degrade performance significantly
- Two options
  - In scripts, can use --wait in between transfers
    - Next transfer only started when previous completes
  - Group transfers together under same Shift id using stdin
    - src(s) and dst, one per line
    - Command line options apply to all lines
    - echo "/file1 /dir1" > in; echo "/file2 /dir2" >> in; shiftc < in
    - Use --hosts to spread grouped transfers across systems

Question? Use the Webex chat facility to ask the Host

A Transfer Is Not Done Until It Is "done"

- Do not remove src files before Shift reports "done"
- Wait for completion email or check status with --status
- Shift may process files in a different order than other tools
  - The existence of a particular file does not imply the existence of others
- Shift may operate on different portions of the same file at the same time
  - A file may show up with full size but still be incomplete

Ignore Errors At Your Peril

- A transfer ending in the "error" state is not complete
  - Files may not have been transferred or partially transferred
  - File integrity may not have been verified
    - Done files may show N/N but some files could be corrupt
  - File sanity checks may not have been completed
  - Tar renames may not have been completed
  - File attributes may be incorrect
    - File attributes are only preserved after its checksum stage
    - Dir attributes are only preserved after all files preserved
- A transfer in the "warn" state has operations that will be retried
  - Errors may be transient and correct themselves after retry

Question? Use the Webex chat facility to ask the Host

Restart Is Your Friend

- Transfers ending in "error" state typically do not need to be rerun from scratch
- Shift provides a restart capability
  - shiftc --restart --id=N
  - Only failed operations will be retried
  - Many transfer options can also be respecified (see man page)
- Some errors are not recoverable
  - Src files that have been removed
  - Dst files that have been removed (mostly tar files but some other cases)
  - Src files that are still being written
  - Shift transfer metadata issues (mostly due to bugs)
- If a restarted transfer keeps failing and the reason is unknown
  - Email support@nas.nasa.gov

Question? Use the Webex chat facility to ask the Host

Avoid Transfers Of Files Still Being Written

- Transfers of unstable files rarely complete due to constant mismatches of src/dst checksums
  - You may start ignoring errors thought to be due to changes
- If necessary, consider use of --include/--exclude
  - For example, changing files always named "bad.file"
    - --exclude='bad.file'

Inclusion/Exclusion

- File names may be explicitly included/excluded
  - Include only files matching given expression (--include)
  - Exclude files not matching given expression (--exclude)
  - May be specified multiple times
- Options take regular expressions – not wildcards
  - --include='foo*' matches 'fo{o,oo,ooo,...}', but not 'foo1'
  - Use '.*' in place of '*' for approximate conversion (e.g. 'foo.*')
  - See "man perlre" for advanced syntax

Question? Use the Webex chat facility to ask the Host

Status Updates

- Status is only updated in between batches
  - Batches with large files take longer in between updates
    - Splitting allows status to be updated more frequently
- Retrieval from lou may require reads from tape
  - Done files can stay at 0/N for hours
  - Done files in tar extractions can stay at 0/0+ for hours
    - File count/size not known until tar file(s) back online
- If you don't think the transfer is doing anything
  - Find relevant host(s) with "shiftc --status --state=run --id=N"
  - Find transfer related processes on host(s) using "ps"
  - cat /proc/pid/io twice in succession to see if I/O being done
  - Shift may be sleeping before retry or waiting for other clients

Question? Use the Webex chat facility to ask the Host

Conclusion

Conclusion

- You've hopefully learned
  - How to perform basic transfer tasks with Shift
  - When and how to use Shift tuning options effectively
  - Some common usage mistakes and how to avoid them
- Usage takeaway
  - Shift simplifies/automates/unifies copy/sync/tar of local/remote data
  - Shift takes care of transfer best practices for you
- Tuning takeaways
  - Use multiple hosts if available and split large files to keep clients busy
  - Avoid overloading a single host
- Mistakes takeaways
  - Install "sup" client in your own account in writable location
  - Wait until transfers "done", don't ignore errors, and recover with --restart

Question? Use the Webex chat facility to ask the Host

Comments, Suggestions, Questions, Issues Welcome

- You can help make Shift better
  - Suggest features that would make your life easier
    - Some current features exist because users asked for them
  - Report bugs and/or other non-optimal behavior
    - User environments and use cases can be very different
    - May not even know a problem exists unless reported
  - Email to support@nas.nasa.gov
- You can even use Shift at your site
  - Available as open source (http://shiftc.sf.net)
    - Latest version will be available in week or two

Question? Use the Webex chat facility to ask the Host